CALIFORNIA STATE BOARD OF EDUCATION

ITEM # 20

JULY 2003 AGENDA

SUBJECT		
Standardized Testing and Reporting (STAR) Program: Adoption of Performance Standards (Levels) for the California Alternate Performance Assessment (CAPA)	X	ACTION
	X	INFORMATION
		PUBLIC HEARING

Recommendation:

Approve for public hearing, and conditionally adopt the Performance Standards (levels) for the California Alternate Performance Assessment (CAPA) for reporting performance levels for 2003.

Summary of Previous State Board of Education Discussion and Action

The California Department of Education (CDE) has provided periodic updates to the State Board of Education on the development and implementation of the California Alternate Performance Assessment (CAPA).

Background. In order to meet the requirements of the Individuals with Disabilities Education Act (IDEA), Title 1, and the new No Child Left Behind (NCLB) Act, the state must show evidence that all students are included in the statewide assessment and accountability systems. The 1997 Amendments to IDEA required all states to develop and implement an *alternate assessment* for children with disabilities who cannot take part in the general statewide assessment programs. Generally this applies to approximately 1% of the total student population. In addition, federal law requires that the results from the alternate assessment be integrated into the state's accountability system.

In response to these requirements, CDE contracted with Educational Testing Service (ETS) to develop and administer the CAPA. The first statewide administration of CAPA took place in Spring 2003, with approximately 45,000 – 50,000 students taking the assessment.

Summary of Key Issue(s)

Federal requirements mandate that alternate assessment results be reported with the same frequency and detail as the general assessments. CAPA Performance Standards must be adopted by SBE in order for this year's CAPA results to be included into the 2003 Base Academic Performance Index (API), and to meet the federal requirement for inclusion in the Adequate Yearly Progress (AYP).

From June 16 – 18, ETS convened a standards setting panel to develop recommendations for Performance Standards based on the Spring 2003 administration of CAPA. Attachment 1, *California Performance Assessment (CAPA) Standard Setting Plan*, describes the standard

Summary of Key Issue(s)

setting process that was used to develop the recommendations. The recommended Performance Standards will be provided in the Supplemental Mailing.

Fiscal Analysis (as appropriate)

None

Attachment(s)

1. *California Performance Assessment (CAPA) Standards Setting Plan* (Pages 1-10)
   The *California Performance Assessment (CAPA) Standards Setting Plan* does not include appendices because they reveal secured test items.

Proposed CAPA Performance Standards will be submitted with the supplemental

California Alternate Performance Assessment (CAPA)
Standard Setting Plan - Revised

Educational Testing Service

Submitted to the California Department of Education

June 13, 2003

Table of Contents

CAPA is an individually administered, standards based assessment for students with moderate to severe disabilities who are unable to take the general STAR assessment with accommodations. CAPA is composed entirely of performance tasks. Each content area includes 8 performance tasks, which are scored by a trained, certificated or licensed school staff member on either a 4 or 5-point rubric depending on the test level being assessed. Currently CAPA is operationally assessing students in the areas of English Language Arts (ELA) and Mathematics with a pretest section in Health. Content areas will be added to the operational administration one per year until the complete test in 2007 contains six content areas: ELA, Math, Health, Physical Education, Science and Social Studies – History. Multiple standard settings will need to be conducted over the next four years to establish cuts in all six content areas.

Goals of the Standard Setting Process

The purpose of the standard setting process is to collect recommendations of the placement of the CAPA cut scores for use by the California Department of Education (CDE). It is imperative that cut score recommendations be based on the knowledge and perspectives of teachers, administrators, parents, and other community members, such as college professors, consultants, or school psychologists, with knowledge of or expertise about this population. The recommendations collected will be presented to the CDE who will have final decision making authority and approval of the cut scores to be used operationally to assign students to the following 5 performance categories: advanced, proficient, basic, below basic, far below basic. The standard setting session will have 3 goals: 1) Set the four cuts necessary for each content area of each test level to enable reporting by the 5 performance categories listed above. 2) Write performance descriptions for the minimally competent student at each of the 4 cut scores. 3) Gather validity related evidence on the items within the assessment and on the assessment as a whole.

Time and Location

A standard setting will be held June 16, 2003 to June 18, 2003 for the areas of ELA and Math at all 5 test levels. The Performance Profile Method, a modified bookmarking procedure, will be used to set standards based on profiles of student test performance. The standard setting session will be held in Sacramento at the Hilton Arden West. Participants will be reimbursed for their travel and lodging pursuant to the California Department of Education guidelines.

The Educational Testing Service (ETS) will secure meeting rooms in which to conduct the standard setting sessions. The California Department of Education (CDE) and ETS will work together to recruit a representative sample of panelists to participate in the standard setting sessions. ETS has created a form to be completed by special needs teachers, administrators and parents to volunteer for the standard setting, as well as the future item writing and content review panel. The goal is to receive enough volunteers for each activity to allow ETS to work with the CDE to select the most representative sample

of volunteers possible. Sixty panelists are being sought for the standard setting with the goal of having a final sample of 45 panelists to divide into 3 groups of 15 members each. Currently, thirty-seven people have been identified for the standard setting. Appendix A lists the data for the volunteers to date. Additional efforts are being made by the CDE to secure panelists from areas of the state, disability groups, and test levels, which need better representation.

Preparation of Standard Setting Materials

Prior to the standard setting session, each panelist will be mailed a letter which explains the purpose of the standard setting, briefly outlines the process that will be followed, their role in the process, and provides a general agenda for the standard setting session. The letter will also explain the security procedures to be followed and discourage the panelists from bringing any personal materials into the standard setting sessions.

On the first day of the standard setting, each panelist will be provided with a copy of the test materials for the test levels for which they are setting standards. Panelists will be assigned an ID number and materials will also include an ID number. A record will be kept of each panelist and the set of materials they receive. Panelists will be required to sign a security agreement, notifying them of the confidentiality of the materials used in the standard setting and prohibiting the removal of the materials from the meeting area. To ensure that all materials are accounted for, the panelist ID and materials ID will be verified at the end of each day and at the conclusion of the standard setting process.

Test items from the Spring 2003 CAPA administration will be presented in the order administered with classical item statistics presented for each item: p+ and polyserial correlation. Along with the test items and statistics, panelists will receive a list of all test items and the content standards which they are intended to measure. For each CAPA content area (ELA and Math), there are five test levels and eight tasks per level. Level 1 uses a 5 point rubric for performance scoring while Levels 2 – 5 use a 4 point rubric. On Level 1 it is possible to obtain any raw score between 0 and 40, on the other test levels raw scores range from 0 to 32.

The basis for the bookmark judgments in the standard setting will be the selection of representative student profiles to represent differential performance across tasks within each content area at each test level. Two to five profiles will be examined at each raw score level. Operational data will be used to ensure that the performance profiles most often achieved are included in the profile packet. Performance profiles representing the same raw score will be grouped in random order since no one item is intended to be more important than another.

The bookmark method is well documented and uses item response data to order test items by difficulty in preparation for the placement of the bookmark (Mitzel, Lewis, Patz, & Green, 2001; Lewis, Green, Mitzel, Baum, & Patz, 1999; Green, Lewis, & Michaels, 2002). One of the advantages to the bookmark method is the relative ease of this method compared to methods, e.g., Angoff or Nedelsky, which require a recorded written response to each individual item. This method also provides rich information to panelists regarding the types of items students perform well on and those that are more difficult. Thus, the cut score takes on more meaning to panelists using a Bookmark

method than when they work with the Modified Angoff procedure. The number of cuts to be set over the three day session in addition to the relative ease of the method is one justification for choosing to bookmark the profile sets in the CAPA standard setting.

Due to the small number of items and the fact that all the CAPA items are constructed response, we anticipated that the typical bookmark method would be confusing for the panelists due to the repetition of each item within the set to represent each point on the rubric. In an attempt to ease the mental burden expected of the panelists, we adopted a modified procedure that takes into account other standard setting research. Donahue, Benson, and Cramer (2000) conducted a standard setting for the Georgia Kindergarten Assessment Program – Revised which combined the judgmental policy-capturing method (Plake & Hambleton, 2001; Jaeger, 1995a; Jaeger, 1995b) and the dominant profile judgment method (Plake & Hamilton, 2001; Plake, Hamilton, & Jaeger. 1997; Putnam, Pence, & Jaeger, 1995). Both of these methods are more holistic than a traditional bookmark method in that they ask panelists to make decisions based on an examinee's score profile or performance rather than on each separate item. Donahue, Benson, and Cramer found that this combination was very successful and teachers reported a high level of satisfaction and confidence with the method used and the resultant cuts. In this standard setting session, each profile was assigned to a performance group and the group assignment was recorded for each profile in a modified Angoff manner. Approximately 50 profiles were used and the process of recording a response for every profile was time consuming.

In order to reduce the mental burden and the amount of time panelists spend on each round of the standard setting process, it was decided that CAPA will use a bookmarking procedure to set multiple cuts concurrently for each test and content level. Using the Performance Profile Method, which is similar to that used in the Donahue, Benson, & Cramer (2000) study, to bookmark the performance profile sets will enable the panelists to set the cut scores in a more holistic manner recognizing the varied strengths and weaknesses encountered in this population of students. An additional advantage is the ability of the panelists to set the cuts in a more efficient manner than occurred when a recorded response was required for every profile.

Representative student profiles will be selected at every raw score. Student performance profiles at the total raw scores of 0, 8, 16, 24, 32, and 40 will not include the combination resulting from the student receiving the same rating for each item (e.g., the score of 16 achieved by the student receiving a score of 2 on all 8 items). These combinations are eliminated to reduce the tendency of the panelists to choose the combination with all 4's or all 3's as the proficiency cut rather than considering the actual student performance across items when placing their bookmarks. For most raw score points, 2 - 3 profiles will be examined but at score points achieved by a large group of students as indicated by the operational data up to 5 profiles may be examined. The most frequently achieved profiles will be chosen for representation at each raw score. Profiles for each test level and content area will be ordered from the lowest total raw score to the highest and placed into a three-ring binder. While it is recognized that any number of combinations of item ratings may result in the same raw score, the intent is to set a cut score that is compensatory in nature. Therefore, profiles within the same raw score will be ordered randomly. Appendix B shows an example of the student profile the panelists will be using in the standard setting process.

Panelists will receive generic performance level descriptions created by the CDE to use as the basis for setting cut scores and for subsequent written descriptions for the minimally competent student at each cut score. These generic descriptions are intended to be used as a starting point for the definition of the minimally competent student at each performance cut. The descriptions are not intended to limit panelists in anyway. Panelists may add or delete any text they desire from the generic descriptions or may start with a blank slate.

A rating form will be provided which lists each profile ordered the same as in the profile binder with three columns, one column for each iteration. (see Appendix C) Additionally, each panelist will receive a survey sheet asking them to rate the importance and relevance of each item and the importance and relevance of the assessment as a whole.(see Appendix D) Finally, the panelist will be asked to provide a rating as to their level of confidence in the standard setting process as a whole.

In addition to the panelists and the measurement professionals who will conduct the standard setting, representatives from the CDE and from ETS Test Development and Program Direction will be on site to hear and when appropriate respond to panelists' questions and concerns during the 3 day session. The Test Developer will also be on hand to review the content standards and answer any questions specific to the content standards or the items.

Process

Beginning on Monday, June 16, 2003, panelists will receive training on the method to be used and will review the generic performance level descriptions created by the CDE. Panelists will also have the opportunity to participate in a mock standard setting for practice before starting the CAPA standard setting. Following the training, panelists will be broken into 3 groups according to the level of student they teach or have experience with to begin the standard setting process. Groups will then be assigned one or more test levels and break out into separate meeting rooms for the remainder of the standard setting session.

Panelists in each group will be asked to review the items in that level, with the items presented in the same format as they are administered. Then panelists will work together to create performance level definitions of the minimally competent student at each cut score for that test level and content area using the performance level definitions created by the CDE as a starting point and referring to the content standards. Once minimally competent definitions have been created, panelists will be asked to independently review the items for the test level and content area.

For each cut score beginning with the cut between basic and proficient, panelists will be asked to begin with the first profile and working toward the back of the set review each profile in the binder. Using the profile binder they will locate the position between profiles where they feel that the minimally proficient student will have a profile this good or better. This is the place where they will insert the bookmark for the cut between basic and proficient. This will be repeated for each cut within that content area of that test level. Panelists will be asked to place bookmarks in the following order: Proficient, Basic, Below Basic, and Advanced. The proficient cut, between proficient and basic, is the first cut to be set because it is the most important distinction. A majority of CAPA

reporting will focus on two groups, those who are proficient or above and those who are below proficient. Therefore, it is most important that this cut be placed in the optimal location, while the remaining cut scores are arranged around the proficient cut.

When all first round cuts have been set, panelists will record the position of their bookmarks on the provided rating sheet (see Appendix B) by drawing a line at the point on the profile list where the cut was made. The selected cut scores assigned to the minimally competent student at each performance level will be recorded by data entry personnel in an Excel spreadsheet. The value recorded is the raw score for the profile immediately following the bookmark. When all panelists have completed the round, the cut scores assigned will be recorded on separate overheads for each cut along with the frequency for that selection if more than one panelist chose the same cut score. This will provide panelists a visual for the variation in cut placement within the group. Beginning with the upper- and lower-most placements for each cut, discussion for each cut will be held as to the rationale for the placements. After discussion, a second iteration using the profile binder will occur where panelists will be asked to revisit their cut score decisions and make any alterations they feel are necessary.

After the 2nd iteration, cuts will once again be displayed to facilitate discussion among panelists along with the median cut score for the total group. Frequency distributions from the operational data will be used to present impact data at the current group cut score. Impact data will be presented for the total student population scored to date, as well as by subgroup for gender, ethnicity, economic disadvantage and disability group. Following large group discussion, panelists will be given the opportunity to alter their cut score decisions one more time. A total group median cut for each cut score will be calculated along with the standard error. The calculation for the standard error will be the standard error of the mean for the judge's opinions. This information along with the impact data will be presented to the CDE following the completion of the standard setting sessions for the final cut score decision.

This process will be repeated within each group over the 3 day period until all cuts are set for each test level and content area. Groups will enjoy a working lunch and break for the day when the cuts for a test level are complete or at an opportune time such as between iterations. Group 1 will be headed by Dr. Lora Monfils and will set cuts for the Level 1 assessment. Group 2 will be headed by Dr. Deanna Morgan and will set cuts for the Levels 2 and 3 assessments. Group 3 will be headed by Dr. Marianne Perie and will set cuts for the Levels 4 and 5 assessments. Group facilitators will be cognizant of the cut scores being set in the other groups and compare cuts by way of impact data frequently to monitor any significant discrepancies that may occur between groups. This will be an indicator of how similar cut scores are across groups in terms of the percentage of students falling into each category based on the 2003 Operational data.

At the end of the last day of the session, panelists will be asked to rate each item in the assessments they worked on as to its importance and its relevance for the students they represent. Panelists will provide the same ratings of importance and relevance for the assessments they worked on as a whole. Finally, panelists will rate their level of confidence in the standard setting approach used and in the cut scores that will come out of this session. Appendix D includes the survey form to collect this information. Participants will also be surveyed about the standard setting session in general and their

contentment level with the food, location and lodging provided. Appendix E includes the survey form for the general session information.

As panelists leave the session on the last day, all materials will remain in the meeting rooms and be 100% accounted for by the standard setting staff. Materials that are no longer needed will be shredded and securely disposed of following the standard setting session.

Analysis of the Data

After each iteration of the standard setting process, the raw score of the profile immediately following each bookmark will be recorded for each cut. Following the first iteration, a graphic of the actual cuts will be presented to facilitate discussion and provide an indication to the group of how the cut score they set compares to that of their group. Then panelists will have an opportunity for a second iteration where they may revisit and revise the placement of the bookmarks from the first iteration.

After the second iteration, the raw score of the profile immediately following each bookmark will be recorded for each cut. We will then take the median of all judgments to find the temporary cut score. This raw score will be located on the cumulative frequency distributions for the Spring 2003 Operational administration and impact data indicating how many students would fall into each performance level based on the second iteration cuts will be presented by total group and subgroups. Following any necessary discussion of the temporary cut score and the resultant impact data, panelists will have a third and final opportunity to revisit and revise their bookmark placements.

Results from the third iteration of the standard setting process will not be released to the panelists as it is still pending approval by the CDE and adoption by the State Board of Education (SBE). Final bookmark placements will be recorded and the median raw score will be located. Bookmark placements for each cut across all panelists in the group will be compiled to determine the median. The standard error of the mean will be computed for each cut as an indicator of variability. This error bands for each cut will be examined for possible overlap due to the relatively small number of raw score points being used with this number of cut scores. Any overlap of error bands will be brought to the attention of the CDE for consideration in the final approval of the cut scores.

Final Cut Decisions

The CDE will approve the cuts and forward the recommendations to the SBE for final adoption of the performance levels to be used operationally. The CDE will be present during the standard setting sessions to hear discussion and observe the process. Cut scores based on the standard setting sessions along with standard errors and impact data for the total group and subgroups will be provided to the CDE on June 19 – 20, 2003. On July 9, 2003 the SBE will be asked to take action on the recommended cut points and to formally adopt the CAPA performance levels for English Language Arts and Mathematics.

Adopted performance levels will enable conversion scoring to begin in preparation for score reporting. Raw cut scores will be matched to the CAPA version of

the STAR scale cut scores at the basic (30) and proficient (35) cuts. All CAPA raw scores will be converted to the 15 – 60 score scale fixing the raw to scale conversions for the basic and proficient cuts. Scale scores will then be converted to one of the five CAPA reporting performance levels.

The final technical report for the standard setting will be produced and delivered to the CDE by July 8, 2003. The technical report will contain a description of the process used to set standards, results from the standard setting process and surveys completed during the standard setting, impact data provided to the CDE, and a listing of all panelists and their qualifications.

References

Donahue, B.H., Benson, J., & Cramer, S. (2000). Standard setting on a kindergarten performance assessment: A joint application of two recent methods. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April 25, 2000.

Green, D.R., Lewis, D.M., & Michaels, H. (2002). Standard setting: A Bookmark Approach. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April 2-4, 2002.

Jaeger, R.M. (1995a). Setting performance standards through two-stage judgmental policy capturing. Applied Measurement in Education, 8, 15-40.

Jaeger, R.M. (1995b). Setting standards for complex performances: An iterative, judgmental policy-capturing strategy. Educational Measurement: Issues and Practice, 14 (4), 16-20.

Lewis, D.M., Green, D.R., Mitzel, H.C., Baum, K., & Patz, R.J. (1999). The bookmark standard setting procedure: Methodology and recent implications. Manuscript under review.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The bookmark procedure: Psychological perspectives. In G.J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives, (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Plake, B. S., & Hamilton, R.K. (2001). The analytic judgment method for setting standards on complex performance assessments. In G.J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives, (pp. 283-312). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Plake, B., Hamilton, R., & Jaeger, R.M. (1997). A new standard setting method for performance assessments: The dominant profile judgment method and some field-test results. Educational and Psychological Measurement, 57, 400-411.

Putnam, S.E., Pence, P. & Jaeger, R.M. (1995). A multi-stage dominant profile method for setting standards on complex performance assessments. Applied Measurement in Education, 8, 57-83.